Natural Language Based Multimodal Interface for UAV Mission Planning

Meghan Chandarana[1], Erica L. Meszaros[2], Anna Trujillo[3], and B. Danette Allen[3]

[1] Carnegie Mellon University, Pittsburgh, Pennslyvania,, USA
mchandar@cmu.edu
[2] University of Chicago, Chicago, Illinois, USA
elmeszaros@uchicago.edu
[3] NASA Langley Research Center, Hampton, Virginia, USA
{a.c.trujillo, danette.allen}@nasa.gov

As the number of viable applications for unmanned aerial vehicle (UAV) systems increases at an exponential rate, interfaces that reduce the reliance on highly skilled engineers and pilots must be developed. Recent work aims to make use of common human communication modalities such as speech and gesture. This paper explores a multimodal natural language interface that uses a combination of speech and gesture input modalities to build complex UAV flight paths by defining trajectory segment primitives. Gesture inputs are used to define the general shape of a segment while speech inputs provide additional geometric information needed to fully characterize a trajectory segment. A user study is conducted in order to evaluate the efficacy of the multimodal interface.

INTRODUCTION

The proliferation of unmanned aerial vehicles (UAVs) has highlighted the need to develop more intuitive human-UAV interfaces for users who may not be expert UAV pilots. Traditionally, human-UAV interfaces are designed for highly skilled users with a domain knowledge of the application and the UAV system itself. These users make use of mouse and keyboard systems to spend hours training and developing mental models of system level nuances (Chen, Wang, & Li, 2009). When given the option, humans naturally choose to interact with UAVs as they would with another person or even a pet (Cauchard, E, Zhai, & Landay, 2015). By mimicking natural human-human communication modalities such as speech and gesture, natural interfaces reduce their dependency on a highly skilled user base and improve their overall system efficiency (Perzanowski, et. al., 2001; Reitsema, Chun, Fong, & Stiles, 2005; Wachs, Kölsch, Stern, & Edan, 2011).

Previous research investigates human-UAV interfaces with vehicles that are collocated with the user. Ng and Sharlin created a gesture interface based on a falconry metaphor (Ng & Sharlin, 2011). In addition, researchers have investigated how intent could be conveyed to robot teammates without defining specific movement sequences for human-robot teams (Nasser, Sturm, & Cremers, 2013; Ende et al., 2011). A speech-based (Quigley, Goodrich, & Beard, 2004) and 3D spatial interfaces (Li et al., 2015) were designed to give users direct control over a UAV's location and/or flight path. In the past, multimodal interfaces that make use of speech and gesture were limited to more traditional graphical user interfaces (Bolt, 1980). More recently, flexible frameworks for direct control of UAV movement allow users to choose a desired input modality/modalities based on their specific application (Suarez Fernandez et al., 2016).

Despite recent research, the usability of multimodal natural language interfaces for UAV mission planning remains unexamined. This paper presents a multimodal natural language interface that combines speech and gesture input modalities. It examines the performance of the multimodal interface in the context of UAV flight path generation. The effect of (1) previous experience with other single input natural language interfaces, (2) previous experience flying a UAV and (3) a user's choice to sit or stand while using the interface on the overall accuracy and user workload are explored.

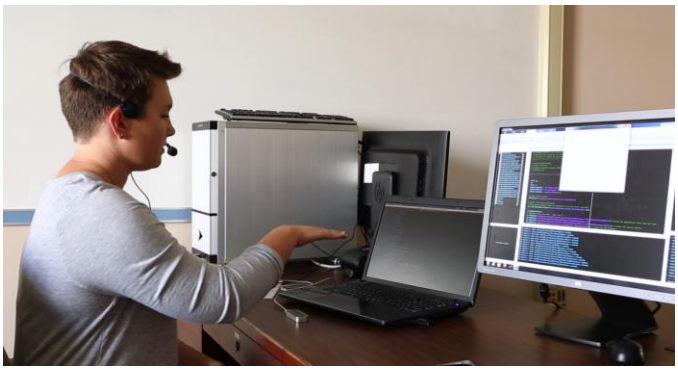

Figure 1: User study setup for the multimodal interface. The user speaks into the headset while performing gestures over the Leap Motion controller.

MULTIMODAL INTERFACE

The experimental, multimodal interface combines speech and gesture inputs to allow users to define trajectory segments in order to build complex UAV flight paths (Fig. 1). Users are able to choose from one of the twelve trajectory segments given in Chandarana et. al.'s library: right, left, forward, backward, left, right, up, down, forward-left, forward-right, backward-left, backward-right, circle and spiral (Chandarana, Trujillo, Shimada, & Allen, 2017). Each trajectory segment's general shape is defined with the gesture module of the interface. Further geometric information – distance, radius, and height – are given using the speech module of the interface. Neither module is individually calibrated for a subject. An interpreter module fuses the speech and gesture inputs such that a fully defined flight path can be generated. Once all desired trajectory

segments have been defined each set of fused data is automatically combined to generate a fully defined flight path for the UAV, which is displayed to the user as visual feedback. The current system instantiation does not allow for changes to be made to the flight path. As in Chandarana et al.'s system, the multimodal interface makes two assumptions about the defined trajectory segments: (1) the *Circle* and *Spiral* segments are defined in the clockwise direction and (2) the *Spiral* segment is defined going upward – height is always a positive change (Chandarana, Trujillo, Shimada, & Allen, 2017).

Speech Module

The Speech Module makes use of CMU Sphinx speech recognition software (Carnegie Mellon University, 2016). CMU Sphinx provides a base English lexicon and mapping of speech sounds to English phonemes that allows for spoken language to be interpreted as text. In order to improve processing time and accuracy, a limited dictionary and grammar were created for this specific speech interface system. The system-specific dictionary contains roughly 100 words corresponding to the geometric information used to define the trajectory segments. The system-specific grammar specifies the order in which the information is expected to appear. This grammar allows for fractional or decimal numbers and different units, and specifies various orders in which the information is expected to occur. For example, units are expected to follow numbers, and directions (height, width, radius) are expected to follow number/unit pairs. As soon as a completed geometrical specification is recognized by the dictionary and grammar, it is immediately sent to the Interpreter Module. Users interact with the speech system using a microphone headset.

Gesture Module

The Gesture Module uses a Leap Motion (Leap) controller (SDK v2.2.6) to track and capture gesture inputs using three infrared cameras. Users make use of $8ft^3$ of hemispherical, interactive space centered on the sensor. During operation the Leap is placed on a flat surface in front of the user such that they could sit or stand depending on their preference.

For each trajectory segment the user wishes to define, they mimic the shape of the trajectory segment with their palm facing the Leap. A Support Vector Machine (SVM) classifier trained by Chandarana et al. is employed to classify the gesture input as one of the twelve shapes in the established library (Chandarana, Trujillo, Shimada, & Allen, 2017). Ten data samples per primitive from 11 users were collected to train the SVM classifier using a linear kernel (120 samples/user). Hand direction movement and the eigenvalues of the hand position throughout the gesture are used as features. The classified shape is then sent to an Interpreter Module. The current model assumes all users are performing gestures with their right hand. After each gesture input, an image of the classified segment is shown to the user as visual feedback. The module then displays a message window which allows a user to either define another trajectory segment by performing the *Right* gesture, or finish and see the total flight path built by performing the *Left* gesture.

Interpreter Module

The Interpreter Module fuses the shape and geometric parameters necessary to define a given trajectory segment by first synchronizing the data given by the speech and gesture modules. In order to fully define a trajectory segment, both the speech and gesture data must be received. However, the different processing times often results in speech and gesture data being received at varying frequencies and in a varying order. In addition, the differences in data types collected must be parsed and integrated. These issues are mitigated by maintaining an individual priority queue of data received from each input module. By preserving the order of data received from each input module, shape and geometric information can be paired based on their place in their respective queues.

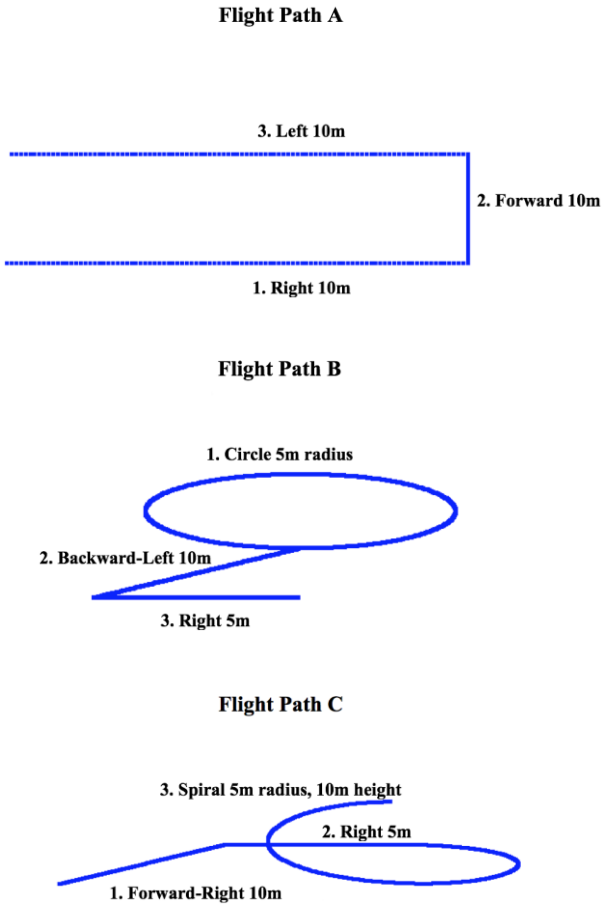

Figure 2: The three flight paths defined by each subject in the user study.

EXPERIMENTAL SETUP

Twelve researchers (some with UI design experience) participated in the user study. Of these twelve subjects four had previous experience using a gesture-based interface for UAV flight path generation and four had previous experience using a speech-based interface for UAV flight path generation, and four had no prior experience. All subjects were either right handed or comfortable using their right hand. Each subject was asked to build three flight paths, which ranged in difficulty level (Fig. 2). The flight paths used were those used by Chandarana et al. (Chandarana, Meszaros, Trujillo, & Allen, 2017) augmented with the geometric parameters needed to fully define the

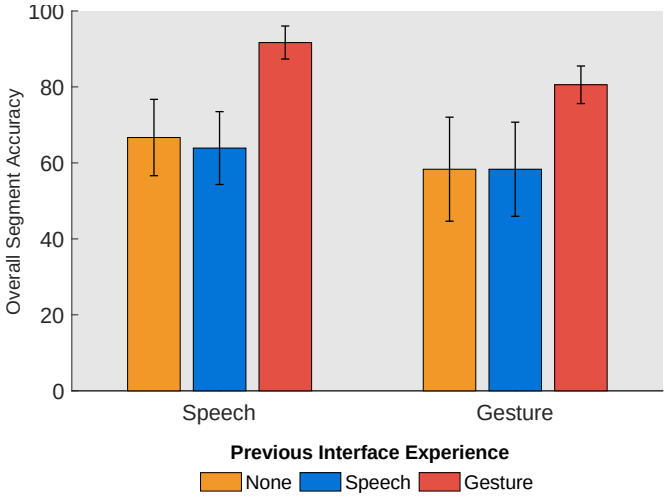

Figure 3: Accuracy of trajectory segment input defined by each module based on subjects' prior interface experience.

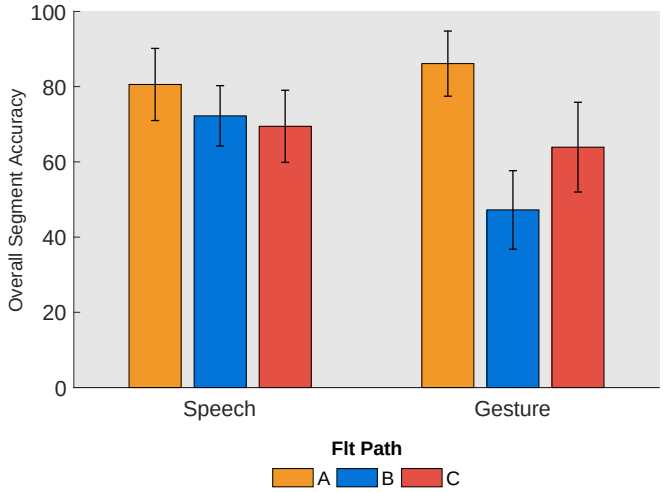

Figure 4: Accuracy of trajectory segment input defined with each module for each flight path.

trajectory segments (i.e., distances, radii, and/or heights). Each flight path contained three segments. A standard *Right* segment was included at different places in the sequence of each flight path to mitigate any biases in segment order.

Before starting the trials, subjects were asked to read and complete a Privacy Act Notice and Informed Consent Form. Next, researchers gave an overview of the user study goals and outlined the general requirements and procedure. Prior to being trained on the interface, subjects filled out a background questionnaire. All subjects were trained on the gesture module first. Once they felt comfortable using the module, the simultaneous input from the speech module was added (e.g., a *Forward* gesture was supplemented with saying "Fly forward 10 meters."). Subjects chose whether to sit or stand. A printout of the trajectory segment library was given to each subject. They could keep the printout during training and the trial runs. The total training time was recorded. Subjects were then asked to build each of the three flight paths. Before each trial a printout of the desired flight path with numbered and annotated segments was given to subjects. They were only allowed to study the flight path for five seconds before starting the trial, but could keep the printout throughout the trial. This reduced the need to memorize the desired flight path. The total time to build the flight path and the correctness of the definition given by each input modality was recorded. Six common types of errors were o when defining trajectory segments with each input module: (1) system misinterpretation – human performed correct gesture, but was incorrectly classified, (2) extra segment added – human defined more than the three required segments, (3) human error – wrong segment or not enough segments defined, (4) a system misinterpretation plus human error, (5) system misinterpretation plus extra segment, and (6) extra segment plus human error. After all trials were completed, each subject filled out a NASA TLX workload assessment survey (Byers, Bittner, & Hill, 1989; Hart & Staveland, 1988).

RESULTS

An analysis of variance (ANOVA) using IBM SPSS version 24 was done on all data collected during the user study.

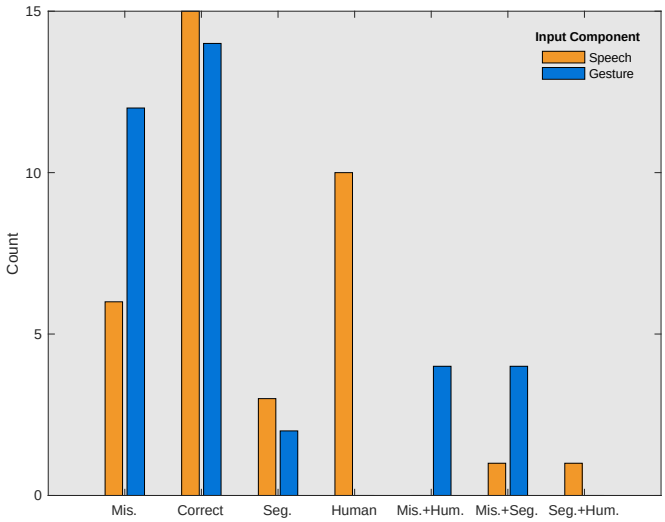

Figure 5: Count of correct and error segments by type when subjects defined speech and gesture components.

Overall subject performance is evaluated given the following independent variables: (1) previous experience with natural language based UAV interfaces, (2) previous experience flying UAVs, (3) flight path, (4) and subjects' choice to sit versus stand while using the multimodal interface. A Tukey HSD Post-Hoc was run on the flight path. Results shown assume a significance level of $p \leq 0.05$. Graphs are shown with error bars for the standard error of mean as appropriate.

All NASA TLX workload measure values given are between 0 and 10. For measures of mental demand, physical demand, temporal demand, effort and frustration a 0 represented low workload, while 10 was high. In performance a 0 indicated that the subject felt they had performed well, while a 10 meant they had done poorly.

The background questionnaire shows that 83.33% of subjects were right hand dominant. However, all subjects were comfortable using their right hand for the trials. 8.33% of subjects had previous experience flying UAVs (RC and/or professional). Their total experience produced an average of 40 hours of flight experience over a 4-year average period. As previously mentioned, one-third of subjects had previous

	Mental	Physical	Temporal	Performance	Effort	Frustration
None	3.88	3.13	3.13	6.13	5.13	4.88
Speech	6.88	4.25	5.25	7.25	6.25	6.25
Gesture	5.13	2.34	4.50	4.25	3.88	3.88

Table 1: Average NASA TLX workload measures given subjects' previous experience with interfaces.

	Mental	Physical	Temporal	Performance	Effort	Frustration
Previous UAV Exp.	4.00	4.00	5.00	5.00	4.00	2.00
No Previous Exp.	5.41	3.18	4.23	5.96	5.18	5.27
Sit	6.10	2.50	4.30	5.90	4.90	5.00
Stand	4.71	3.79	4.29	5.86	5.21	5.00

Table 2: Average NASA TLX workload measures given subjects' previous experience flying UAVs and their choice to sit versus stand.

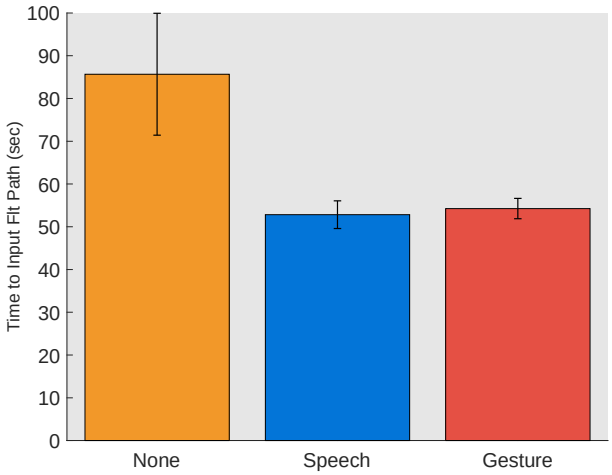

Figure 6: The average time to define flight paths given subjects' previous interface experience.

experience with a gesture-based UAV interface, one-third had previous experience with a speech-based UAV interface, and one-third had no prior experience with a natural language based UAV interface.

Accuracy

Subjects with previous gesture interface experience were more accurate in defining both the speech and gesture components (Fig. 3). Both components of flight path A were more accurately defined (Fig. 4). For the speech components flight path C was the hardest to define, while subjects had the most difficulty with defining the gesture components of flight path B. The gesture component accuracies were statistically significant ($F_{(2,24)}=3.586$ *and* $p=0.043$). The accuracy of the gesture component of flight paths A and B were statistically different. Fig. 5 shows that the gesture component given by subjects was misinterpreted more often. A greater number of human errors was seen when defining the speech component.

Input Time

Fig. 6 shows that subjects who had no previous experience with a natural language based UAV interface took the most amount of time to define the flight paths. Those with previous speech interface experience took slightly less time than those with prior gesture interface experience. The Fig. 6 results were significant with $F_{(2,24)}=3.702$ and $p=0.04$. Flight path C took the least amount of time on average to build followed by flight A and then B (*56.92 sec*, *67.00 sec*, and *68.83 sec*, respectively). Subjects with no previous UAV flight experience took longer to build flight paths than those who did (*58.00 sec* and *64.82 sec* respectively). Users who chose to stand took less time to input flight paths than those who chose to sit (*59.81 sec* and *70.47 sec* respectively). The input time was negatively correlated with training time for subjects with previous gesture interface experience, but positively correlated for those who had previous speech experience or none at all (Fig. 7).

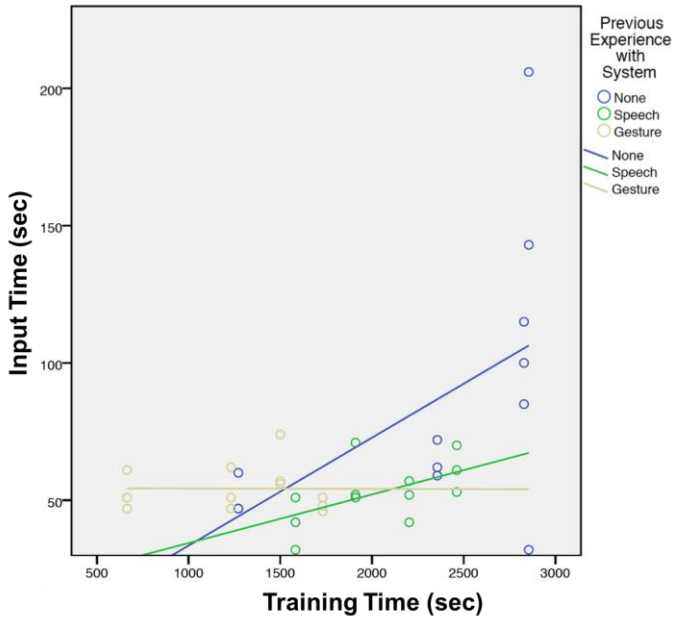

Figure 7: Correlations between training time and average time to input flight paths given subjects' previous experience with interfaces

Subjective Measures

Tab. 1 shows the average NASA TLX workload ratings given by subjects after using the multimodal interface. The results are separated by subjects' previous experience using natural language interfaces. Those with previous experience with speech interfaces rated their workload the highest. In all measures except for mental and temporal demand subjects with previous gesture interface experience had the lowest workload ratings. Tab. 2 shows that subjects who had previous experience

flying UAVs rated their workload lower than those who did not for all measures except for physical and temporal demand. The choice to sit versus stand had little effect on subjects' temporal demand, performance, effort and frustration. Standing produced a lower mental demand, but higher physical demand.

DISCUSSION

Although a small subject sample size was used for this initial evaluation, we observe that the relatively high accuracy of subjects with no prior interface experience and less than an hour of training time indicates that the multimodal interface was fairly intuitive to learn (Fig. 3). The lower general workload measures for subjects with gesture experience (Tab. 1) indicates that their prior experience with a similar highly spatial interface allowed them to learn the multimodal interface more effectively and easily than other subjects. Therefore, subjects who had previous experience with a gesture interface were able to define trajectory segments more accurately than subjects who had previous speech interface experience or no experience at all (Fig. 3). Surprisingly, these subjects were even able to define speech components better than those with previous speech interface experience. This also resulted in a negative correlation between input time and training time as compared to subjects without gesture experience (Fig. 7). Although subjects with no experience took longer to build the flight paths (Fig. 6), they felt less workload in general than subjects who had experience with speech interfaces (Tab. 1). This suggests that the gesture input module was easier to learn how to use when there was no expectation of how a similar interface should work.

Different flight paths gave subjects difficulty when defining speech and gesture components (Fig. 4). The speech component of flight path B was easier to define than the gesture component and vice versa for flight path C. Since both flight paths contained a straight and diagonal trajectory segment, the difference can be attributed to the difference between defining a *Circle* and *Spiral* segment. Overall, with less than an hour of training time, subjects had fairly good accuracies, indicating that it was intuitive to learn.

The familiarity of UAV capabilities from previous flight experience resulted in a lower mental demand, feeling of effort and frustration (Tab. 2). This is evident in their overall lower average time to build flight paths. Standing also helped subjects input flight paths faster. This resulted in a higher physical demand, but lower mental demand. Sitting and standing were equally frustrating. This along with the close ratings for temporal demand, performance and effort indicate that although there was a difference seen in the time to input flight paths subjects did not feel the difference.

CONCLUSION

This paper presented a multimodal interface which used a combination of both speech and gesture inputs to define a complete UAV flight path. Gesture input is used to define the shape while speech is used to provide additional geometric information. The results show that the interface was intuitive. The gesture module was harder to learn how to use than the speech module. Future multimodal interfaces will need to focus on the integration of the input modalities in order to improve overall accuracy rather than forcing users to interact with the interface in a particular way (e.g., standing instead of sitting).

ACKNOWLEDGEMENTS

The authors would like to thank Jeremy Lim of Pennsylvania State University for initial development of the interpreter module and Javier Puig-Navarro of the University of Illinois Urbana-Champaign for fully creating the combined flight path.

REFERENCES

Bolt, R. A. (1980). *"Put-that-there": Voice and gesture at the graphics interface* (Vol. 14, No. 3, pp. 262-270). ACM.

Byers, J. C., Bittner, A. C., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary. *Advances in industrial ergonomics and safety I*, 481-485.

Carnegie Mellon University, "Cmu sphinx4-5prealpha," 2016, retrieved: Jan. 2016. [Online]. Available: http://cmusphinx.sourceforge.net/.

Cauchard, J. R., Zhai, K. Y., & Landay, J. A. (2015, September). Drone & me: an exploration into natural human-drone interaction. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 361-365). ACM.

Chandarana, M., Meszaros, E., Trujillo, A., Allen, B.D. (2017). Fly like this: Natural Language Interfaces for UAV Mission Planning. In: *Proceedings of the 10th International Conference on Advances in Computer-Human Interaction*. ThinkMind (*in press*)**.**

Chandarana, M., Trujillo, A., Shimada, K., & Allen, B. D. (2017). A Natural Interaction Interface for UAVs Using Intuitive Gesture Recognition. In *Advances in Human Factors in Robots and Unmanned Systems* (pp. 387-398). Springer International Publishing.

Chen, H., Wang, X. M., & Li, Y. (2009, November). A survey of autonomous control for UAV. In *Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on* (Vol. 2, pp. 267-271). IEEE.

Ende, T., Haddadin, S., Parusel, S., Wüsthoff, T., Hassenzahl, M., & Albu-Schäffer, A. (2011, September). A human-centered approach to robot gesture based communication within collaborative working processes. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on* (pp. 3367-3374). IEEE.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, *52*, 139-183.

Li, N., Cartwright, S., Shekhar Nittala, A., Sharlin, E., & Costa Sousa, M. (2015, October). Flying Frustum: A Spatial Interface for Enhancing Human-UAV Awareness. In *Proceedings of the 3rd International Conference on Human-Agent Interaction* (pp. 27-31). ACM.

Naseer, T., Sturm, J., & Cremers, D. (2013, November). Followme: Person following and gesture recognition with a quadrocopter. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on* (pp. 624-630). IEEE.

Ng, W. S., & Sharlin, E. (2011, July). Collocated interaction with flying robots. In *RO-MAN, 2011 IEEE* (pp. 143-149). IEEE.

Perzanowski, D., Schultz, A. C., Adams, W., Marsh, E., & Bugajska, M. (2001). Building a multimodal human-robot interface. *IEEE intelligent systems*, *16*(1), 16-21.

Quigley, M., Goodrich, M. A., & Beard, R. W. (2004, September). Semi-autonomous human-UAV interfaces for fixed-wing mini-UAVs. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on* (Vol. 3, pp. 2457-2462). IEEE.

Reitsema, J., Chun, W., Fong, T., & Stiles, R. (2005). Team-centered virtual interactive presence for adjustable autonomy. In *Space 2005* (p. 6606).

Suarez Fernandez, R., Sanchez Lopez, J. L., Sampedro, C., Bavle, H., Molina, M., & Campoy Cervera, P. (2016, June). Natural user interfaces for human-drone multi-modal interaction. In *Proceedings of 2016 International Conference on Unmanned Aircraft Systems (ICUAS)*.

Wachs, J. P., Kölsch, M., Stern, H., & Edan, Y. (2011). Vision-based hand-gesture applications. *Communications of the ACM*, *54*(2), 60-71.